UNIVERSITY OF TORONTO

ExArch

WP3

EXASCALE CLIMATE DATA ANALYSIS

From the Inside Out

Frédéric Laliberté

Paul Kushner

University of Toronto

UNIVERSITY OF TORONTO

ExArch

WP3

A USER PERSPECTIVE

EXASCALE CLIMATE DATA ANALYSIS

From the Inside Out

Frédéric Laliberté

Paul Kushner

University of Toronto

ExArch Work Package 3

- Quality Assurance and Climate Science Diagnostics with CMIP5/CORDEX data archive.
- **U of Toronto**, DKRZ and UCLA.
- Develops a library of climate diagnostics parallel to the development of the ExArch query system.

Challenges

For Users

* Diagnostics using high-frequency data require large downloads, using up bandwidth and demanding vast storage capabilities
* Diagnostics used in intercomparison Studies increasingly require numerically consistent data

For Modeling groups

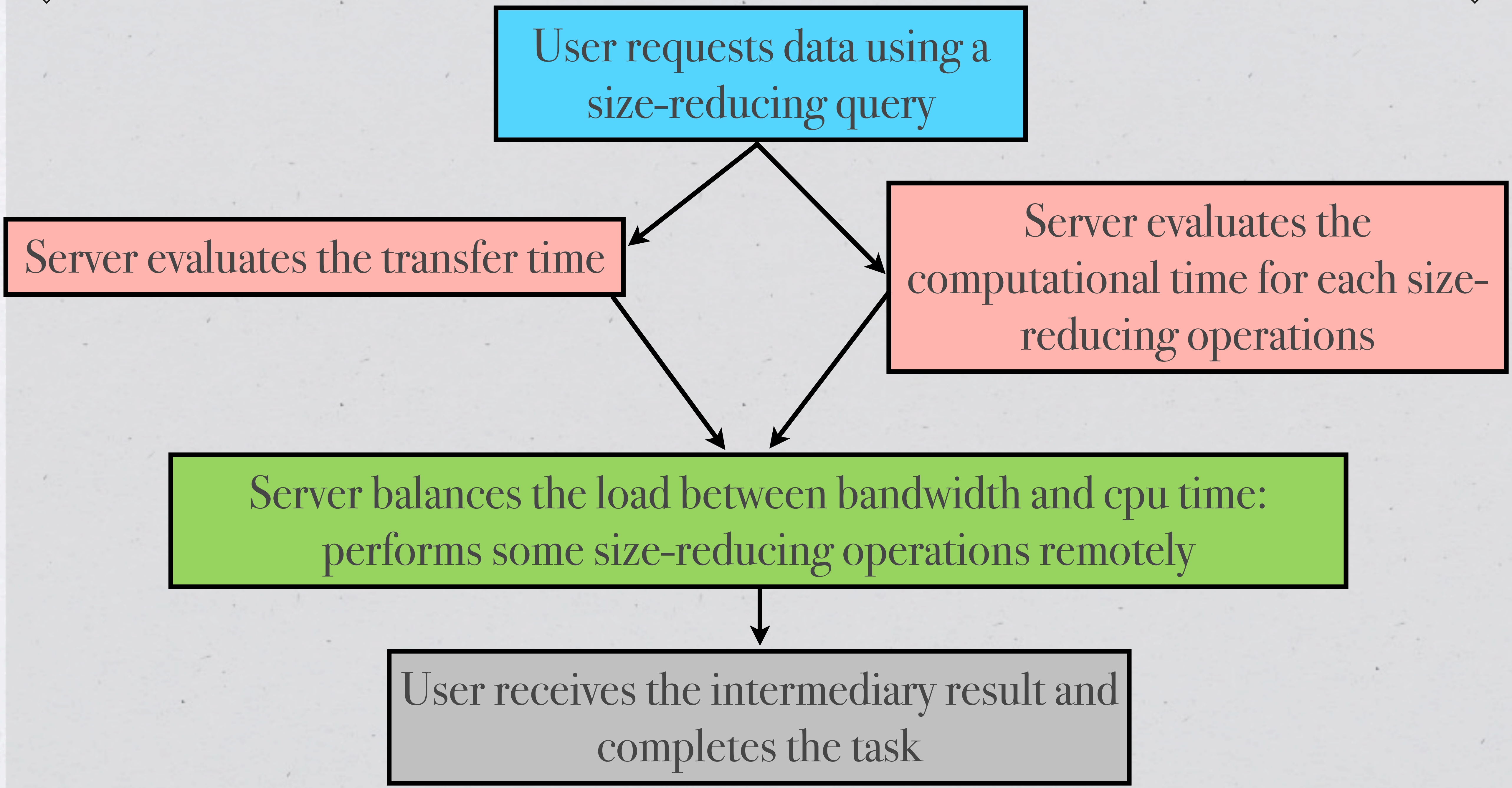
* Data converted from the model's hybrid grid to pressure grid can lead to numerical errors with some diagnostics (e.g. derivatives)
* These errors degrade the model's perceived performance and obfuscates explicit conservation laws

USER PERSPECTIVE

ExArch WP3 - U of T

- Will try to answer the users requirements by benchmarking a series of advanced climate diagnostic using a simple server-side processing framework.
- Expect to better monitor the role that can be played by OPeNDAP in conjunction with CDOs.
- Will prepare advanced climate diagnostics scripts to be ready for to monitor the performance of the tools developed by ExArch associated teams.

Server-Side Processing: An Ideal Case



Server-Side Processing: The reality

User requests data using a size-reducing query

Query requests load balancing estimates from the user

Server estimates cpu of server-side operations and generates a local queue

If cpu queue is too long

User receives the intermediary result and completes the task

Case Study: Isentropic coordinates

For many diagnostics, like the overturning circulation, it is important to compute zonal and temporal averages along isentropic surfaces:

$$\Psi_\theta(\phi,\theta) = \int_0^\theta 2\pi a[\rho_\theta v(\lambda,\phi,\theta',t)]d\theta'$$

	Grid	Size / year at N80
Input	6 hourly data in lat-lon-hybrid	160 x 320 x 60 x1500 x #years
Output	lat-theta	160 x 128

A reduction in size by a factor 200 000 per year!

Isentropic coordinates

Because zonal and temporal averaging are cheap,
the server should compute them:

(1) Convert the meridional velocity and pressure to isentropic coordinates

(2) Compute the isentropic layer mass $\rho_\theta = \frac{1}{g}\partial_\theta p$

(3) Compute the zonal and temporal average of the product $[\rho_\theta v]$

For the user, this process is more transparent and does not require the handling of vast amounts of data.

For the server, bandwidth usage is greatly reduced.

Case Study: Joint Distribution

The mass flux joint distribution can be written as:

$$M(\phi,\theta,\theta_e) = a\cos\phi \frac{1}{T}\int_0^T \int_0^{2\pi} \rho_\theta v(\lambda,\phi,\theta,t)\delta\big(\theta_e(\lambda,\phi,\theta,t)-\theta_e\big)d\lambda dt$$

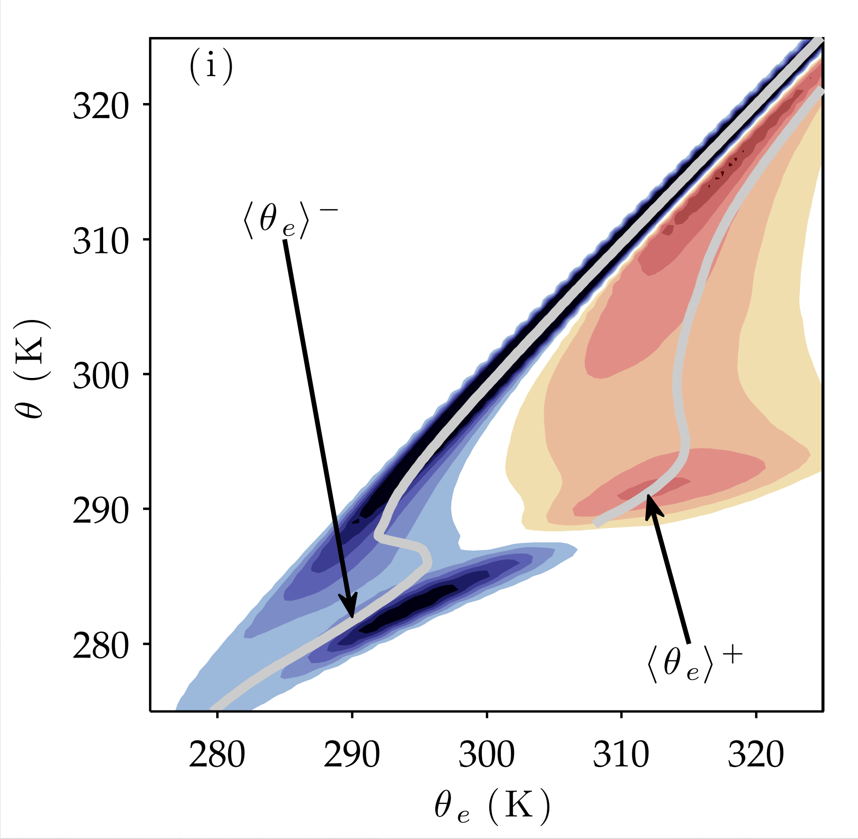
It computes the distribution of equivalent potential temperature on isentropic surfaces

	Grid	Size / year at N80
Input	6 hourly data in lat-lon-hybrid	160 x 320 x 60 x 1500 x #years
Output	lat-theta-equivalent theta	160 x 128 x 128

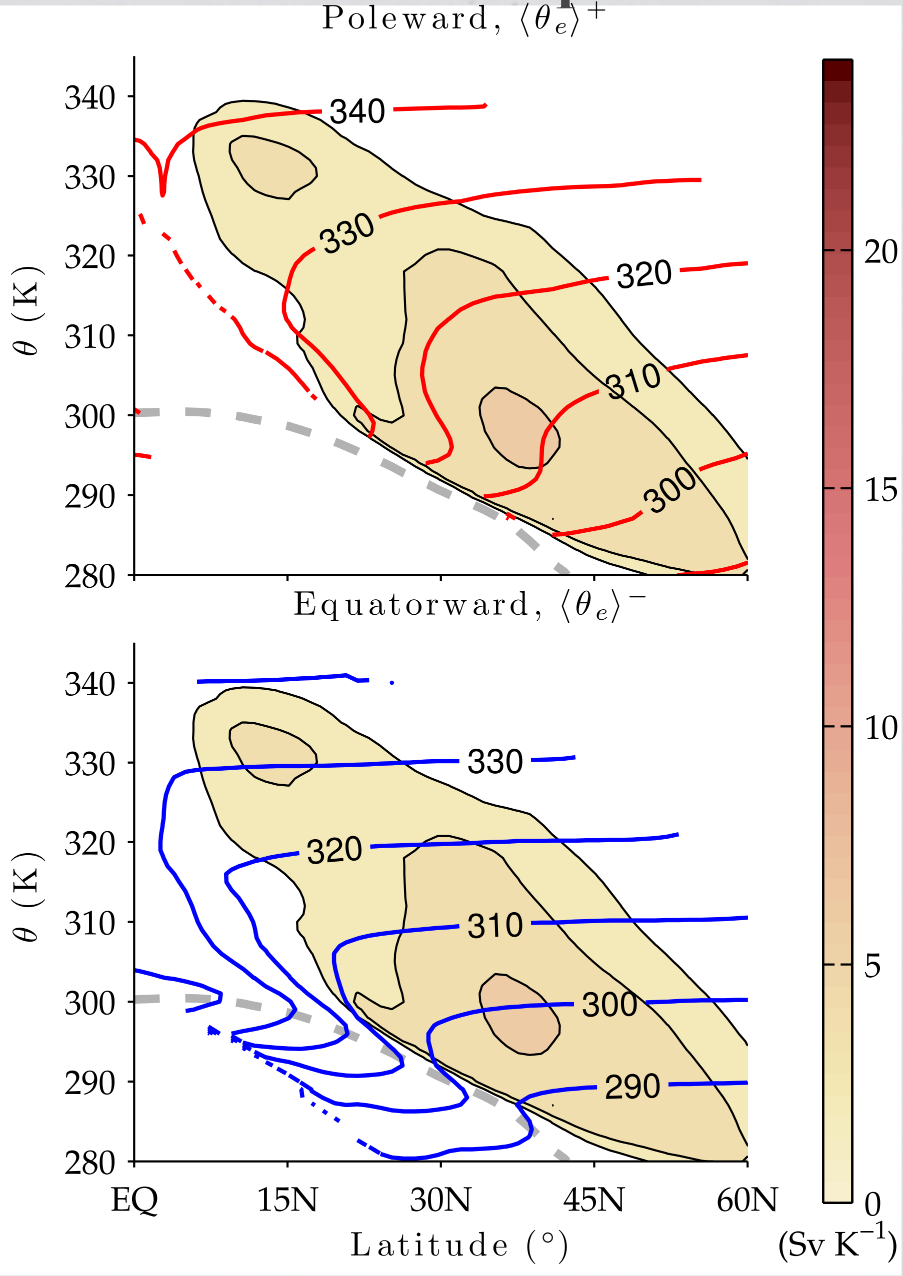
A reduction in size by a factor 1 000 per year!

The Joint Distribution can be used to determine *Lagrangian trajectories of moist flows*

Joint distribution at 35N for DJF



Average equiv. pot. temp in directional components



Other Diagnostics

- Many other diagnostics are based on EOFs (e.g. NAO) and thus require long time-series that are reduced into a few principal components.

- Tropical diagnostics of intraseasonal variability relies on the analysis of space-time spectra (Wheeler and Kiladis, 1999). Newer diagnostics (Dias 2010, thesis) use average over an ITCZ-following latitudinal region.

- Both methods require time series over large area and have huge data burdens to produce simplified diagnostics.

MODELING GROUPS PERSPECTIVE

Functional Data Structure

For theoretical studies of model outputs, it is convenient for the user to consider the data as ‘functions’:

The data files specifies how mathematical operators should combine its variables to be consistent with the numerics.

Take, for example, the computation of the Moist Static Energy (MSE):

$$m = c_p^* T + L_v q_v$$

Depend on moisture parameterizations and are thus model-dependent

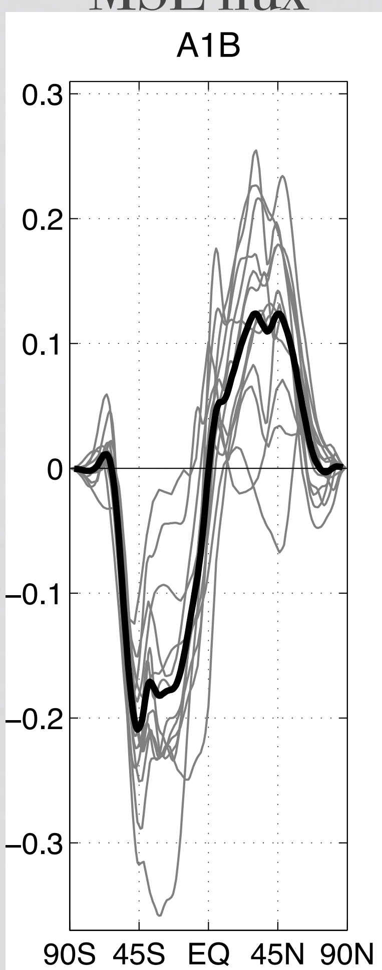
The MSE flux divergence is an important quantity since it is constrained by evaporation:

$$\nabla \cdot \int_0^{p_s} m\mathbf{v} dp \approx E - C$$

This integral will be accurate only if the divergence mimics the conservation law

Next Generation Diagnostics

Zonal Mean
MSE flux



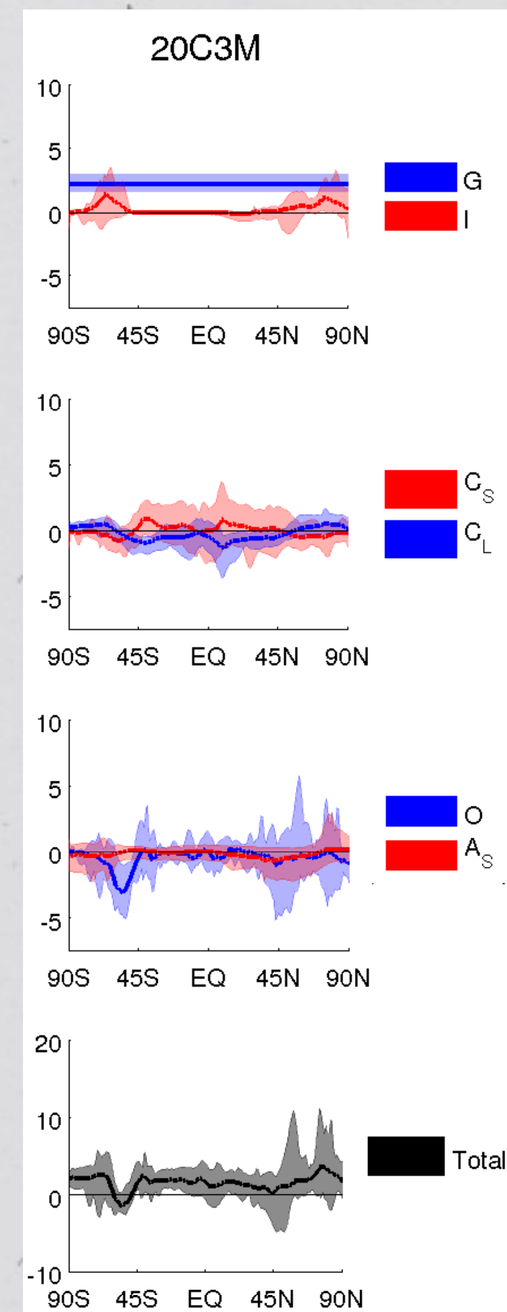
Hwang & Frierson
(2010)

Increasingly, diagnostics will want to consider the data in <u>simplified dynamics models</u> where explicit conservation properties will be important.

Using an EBM for the attribution of climate change components impacts:

$$S - (L_S - L_C) = -\frac{p_s}{g} D \nabla^2 m + F_s,$$

Requires the inversion of the Laplacian using model data.



Conclusion

- With server-side processing, modeling groups would make the development of <u>advanced diagnostics easier and their computation more timely</u>.
- Providing derived data computed from a native grid would <u>reduce numerical errors and improve model intercomparison</u>.
- Modeling groups would make sure that their model performs in the way intended, <u>reducing model bias due to different grid specification</u>. This is particularly important for CORDEX experiments.
- It is hoped that these advantages are worth the <u>extra programming overhead</u> for the modeling groups.